DOCUMENT RESUME

ED 423 306 TM 029 114

AUTHOR Stansfield, Charles W.; Kahl, Stuart R.
TITLE Lessons Learned from a Tryout of Spanish and English Versions of a State Assessment.
PUB DATE 1998-04-09
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Change; Elementary School Students; Elementary School Teachers; Elementary Secondary Education; English; *Field Tests; High School Students; Hispanic Americans; Secondary School Teachers; *Spanish; State Legislation; *Test Construction; Test Format; Test Items; Testing Programs; Translation
IDENTIFIERS *Massachusetts Comprehensive Assessment System; Reform Efforts

ABSTRACT

The Massachusetts Comprehensive Assessment System (MCAS) is the new Massachusetts state assessment program that is being implemented in response to state education reform legislation. The paper describes the early efforts of the state Department of Education (MDOE), its prime contractor for development of the MCAS (Advanced Systems in Measurement and Evaluation (ASME)), and its subcontractor (Second Language Testing, Inc. (SLTI)) in developing the Spanish language version of the MCAS. The paper documents the procedures followed, examines the data collected, and reports informally on what has been learned from the experience. To begin with, since the examinees came from different Hispanic backgrounds, it was decided to use standard Spanish in the examination with certain dialectical variants of words as a gloss in brackets as needed. The items in the 1997 Spanish tryout were distributed across many English forms, so that no one English form corresponded to the Spanish form. After translation of the selected items, an iterative procedure of draft, review, and revision of the forward translation was used instead of back translation as a quality control procedure. Sixteen steps in the adaptation process are listed. Another issue was the format of the test booklets. It was decided to produce the Spanish version in a Spanish-only and a Spanish/English (on facing pages) version. After the Spanish test was administered, interviews were conducted with 97 students in grades 4, 8, and 10 at 19 schools. Seventeen teachers were interviewed after they administered the tests. Scoring was assessed and test items were analyzed. Although it is difficult to draw firm conclusions from the tryout data, the students who took it would not have been able to participate in the regular English test, and so received a benefit from the tryout version. These early results indicate that the test will help address the assessment needs of students, teachers, and parents. An attachment contains the mathematics bilingual version for grade 4. (Contains eight tables and seven references.) (SLD)

ED 423 306

LESSONS LEARNED FROM A TRYOUT OF SPANISH AND ENGLISH VERSIONS OF A STATE ASSESSMENT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Charles W. Stansfield
Second Language Testing, Inc.
10704 Mist Haven Terrace
N. Bethesda, Maryland 20852
ph. 301-231-6046

and

Stuart R. Kahl
Advanced Systems in Measurement
and Evaluation, Inc.
171 Watson Road
Dover, NH 03820
ph. 800-431-8901

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Charles Stansfield
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

TM029114

Paper presented at a Symposium on Multilingual Versions of Tests at the Annual Meeting of the American Educational Research Association, San Diego, CA, April 1998

April 9, 1998

The Massachusetts Comprehensive Assessment System (MCAS) is the new Massachusetts state assessment program that is now being implemented in response to the Massachusetts Education Reform Law of 1993. This law creates a states testing program based on state-approved content standards in each subject. The law requires that the state's testing program:

1. be administered annually in at least grades 4, 8, and 10;
2. measure performance on the academic learning standards contained in the Massachusetts curriculum frameworks;
3. report annually on the performance of individual students, schools, and districts;
4. serve as one basis for a system of student, school, and district-wide accountability;
5. include the participation of virtually all students enrolled in the Commonwealth's public school system, including students with special needs or with limited English proficiency;
6. eventually establish a performance level on the grade 10 test as a high school graduation requirement (MDOE, 1998).

It is the intent of the program to have all students who can meaningfully respond to the MCAS instruments participate in testing. In May of 1998, Massachusetts's students in grades 4, 8, and 10 will be tested in the English Language Arts, Mathematics, and Science and Technology. A test of History and Social Science will be added in the spring of 1999. Consequently, an item tryout for the History and Social Science tests will be a part of the May 1998 test administration. A foreign language test will be added a year later, in accordance with a 1994 amendment to the Education Reform Law. While state policies on accommodations are still being established, it already has been determined that limited English proficient (LEP) students must take the tests if they have attended school in the continental United States for more than three years. The Massachusetts Department of Education has also determined that Spanish-speaking LEP students who have been in the United States for three years or fewer may take a Spanish version of the Mathematics and Science and Technology tests. Spanish speakers compose the largest group of LEP students in the state, just as they do in the nation. Nationally, 73% of all LEP students are native speakers of Spanish (August & McArthur, 1966, cited in Lachat, p. 77).

This paper describes the early efforts of the Massachusetts State Department of Education (MDOE), its prime contractor for development of the MCAS, Advanced Systems in Measurement and Evaluation (ASME), and it subcontractor, Second Language Testing, Inc. (SLTI), in developing the Spanish language version of the MCAS. The purpose of the paper is to document the procedures followed, to examine the data collected, and to report informally on what we learned from the experience.

The Design of MCAS Assessments

MCAS consists of standards-based assessments being developed by the Department of Education in collaboration with committees of Massachusetts teachers and Advanced Systems in Measurement and Evaluation, Inc. The development committees have been meeting since January 1996 to develop test items that address the learning goals identified in the state's

curriculum frameworks. Customized, state-developed assessments are being used for several reasons. First, such assessments are more sensitive to the changes in instruction called for by the state's curriculum frameworks. Second, a significant portion of the items will be released to the public along with the test results each year. These released items will benefit parents and teachers considerably in interpreting the results, and they can be used in other ways, such as for local instruction and assessment, and for preparing students for future MCAS tests. Another widely perceived advantage of customized assessment is that educators across the state, who are involved in designing the program and developing the tests, become more supportive of it. Such support of the assessment program by the educational community may be critical to the success of the MCAS and the state content standards on which it is based.

The MCAS assessments include multiple-choice items and open-response items. Each open-response item requires eight to ten minutes of response time and a half page to a full page of response space. These open-response items are scored 0, 1, 2, 3, or 4, instead of 0 or 1, as multiple-choice items are scored. The Mathematics tests also includes short-answer items, also scored 0 or 1. Non-multiple-choice items count for approximately one-half of the total score on the tests.

Most of the items on each subject test will be the same for all students. These "common" items are the ones that will be released to the public with the test scores. Each student will also take a few additional items of each type that are unique to his or her test form. Twelve different forms will be used in each grade for each subject. Performance on these "matrix-sampled" items will count toward the test score, but these items will remain secure. From this set of secure items, the "common" items for the following year will be selected. Ongoing test item development will replace matrix-sampled items passing through this system.

The results of the tests will be used to assist local educators in improving teaching and learning, to provide for school and district accountability, and ultimately to provide for student accountability. At grade 10, some passing score on the test is scheduled to be tied to high school graduation beginning with the class of 2003. Student level results will include performance level designations in each subject area, as well as normative scores, and item level results. At the school level, percentages of students at four performance levels will be reported, as will scores on the subject area tests and subtests and item level results. Scores on the subject area tests for subgroups of student populations based on gender, ethnicity, school programs, classification as an LEP or special needs student, etc., will be reported at the district and state levels.

The Spanish Tryout

Lacelle-Peterson and Rivera (1994) have called for field testing state assessments on LEP students, in order to help ensure that the instruments are valid for this subpopulation of examinees. Field testing would include LEP students taking the English version of the test and LEP students taking any non-English versions. However, the seven states that currently offer non-English versions of state assessments do not always field test those versions because the results of field testing in two languages might suggest different, even conflicting changes in the

test items. As a result, states generally field test in English and then adapt the final form of the English test to any non-English languages in which it has been decided to test.

Item tryouts for the Mathematics and Science and Technology assessments were conducted statewide in the spring of 1997. (Item tryouts for the English Language Arts assessment were conducted in the fall of 1997.) Since Spanish versions were to be developed ultimately for the operational testing in Mathematics and Science and Technology, the MDOE decided to include Spanish speaking students in the item tryouts as well. The tryout of the Spanish items was designed as a simulation of operational testing, rather than as a traditional field testing, in order to gain experience that could be applied to the first operational versions in Spanish. The Spanish tryout was also designed to emphasize to schools the intent to accommodate Spanish speaking students in the state testing program.

The item tryout forms for the general population were constructed differently from the single form administered to Spanish-speaking students at each grade level in each subject. Each English form consisted of a homogeneous set of items in terms of content. Thus, the forms tended to include items measuring related content. This assured the highest quality data for purposes of item analysis. The item tryout forms were considerably shorter than the forms to be used in operational testing -- 20 multiple-choice and four open-response items. Large numbers of English forms were administered using this matrix sampling design, again simply to generate items statistics for purposes of analysis and improvement of items. The Spanish form was not designed to accomplish this purpose; consequently a more heterogeneous set of items, in terms of content, was selected for the Spanish form. The items to be included in the spring 1998 operational Spanish form were not determined until long after the tryout of items on the English forms. The items in the 1997 Spanish tryout were distributed across many English forms; thus, no one English form corresponded to the Spanish form.

Adaptation to Spanish

In adapting a test to another language, a number of decisions have to be made. Depending on the nature of the original test, on the target language, and the intended examinee population, the adapted test may be very similar to or quite different from the original. In this case, because of the nature of the subjects being tested (math and science), and their link to the state standards, it was agreed ahead of time that the basic content of the tests should remain the same, if possible. Since the intended examinees were known to come from different Hispanic countries, representing a variety of dialects rather than a single dialect, it was decided to use standard Spanish in the test, and to include certain dialectal variants of words as a gloss in brackets as needed.

Brislin (1970; 1976; 1986; Brislin, Lonner, & Thorndike, 1973) has written extensively about back translation, but) translation as a method that produces a high quality translation of a test instrument. A number of other authors (Warner and Campbell, 1969; Bernard, 1988; Hambleton, 1994; McKay et al., 1996) have written about it as well. Back translation as described in the literature on cross-cultural research, involves asking a bilingual to translate the original test to the target language, and then having a different bilingual translated it back to

English. The two English versions are then compared, and points of disagreement are used to identify problems in the initial forward translation. The forward translation is then corrected. Back translation is viewed in the literature as a method for drafting, reviewing, and revising a translation. Actually the purpose of the back translation is to identify and correct errors in the forward translation.

In the opinion of the test translation contractor, there are a number of flaws in relying on back translation to examine the quality of the translated document. First, the lack of agreement between the original document and the back translation may be due to problems with the back translation, not to problems with the forward translation. That is, the back translation is as likely to contain translation errors or infelicities as is the forward translation. Errors in back translation are just as common as errors in forward translation. One is left with two translations and no verification of the quality of either. Once the back translation is completed, the focus of work becomes a comparison of the two English documents (the original test and the back translation), as one searches for points of inconsistency. Next one must search the two translations to determine the reason for the inconsistency. If the reason relates to the forward translation, only then does one begin to consider the solution to problem.

Second, when a translator knows that the initial forward translations will be checked by a back translation procedure, this influences the nature of the forward translation. By producing a very literal forward translation, the translator can ensure that the back translation will produce a document that is highly similar or identical to the source document. However, the literal forward translation may represent stilted rather that natural expression in the target language. As a quality control procedure, back translation essentially ignores such stilted language, even though it may make the translation difficult to read.

Third, because back translation can sends false messages about the forward translation, and because it encourages a literal and unnatural forward translation, it can result in a considerable waste of time and money. The money expended on back translation can be better spent on other aspects of the testing program.

Fourth, the literature on back translation relies on the use of bilinguals rather than professional translators to do the translation. Professional translators, who are themselves bilingual, normally have outstanding skills in written expression in the target language. Selecting someone to do a translation merely because they claim to be bilingual is not likely to produce a high quality equivalent document in the target language.

For the above reasons, it was decided not to use back translation as a quality control procedure for the initial item tryout. Instead, an iterative procedure consisting of draft, review, and revision of the forward translation, was used. The 16 steps set up by SLTI to adapt the MCAS tests to Spanish follow. These steps have been described elsewhere (Stansfield and Auchter, 1997) and are not discussed here.

1. Review test to identify culturally loaded items.
2. Identify professional translators with appropriate background and experience.
3. Identify reviewers with appropriate background and experience.

4. Provide orientation and information on test to translators.
5. Translators produce first drafts.
6. Reviewers review drafts.
7. Revise drafts based on reviewers' comments.
8. Present drafts to Massachusetts Department of Education (MDOE).
9. MDOE conduct review of communities.
10. Send suggested revisions for each test to SLTI.
11. SLTI review suggested revisions and forward to translators.
12. Translators revise, based on MDOE suggested revisions.
13. Identify and describe points of disagreement and send to MDOE.
14. Resolve points of disagreement.
15. Prepare and proof final drafts.
16. Send final copy to test publisher.

A preliminary review of the instruments by SLTI showed that only two items needed to be replaced with items from other test forms in English. This may be due to the fact that the instruments had been extensively reviewed according to standard criteria for fairness and sensitivity. The items identified in the SLTI review involved assumed knowledge of American culture. For example, one assumed a knowledge of how American football is played. Another change that was made in the instruments involved translating English names to Spanish (James ⇒ Jaime), so long as the names were easily translatable.

Two educated native speakers of Spanish were identified to translate the tests. Each was a professional translator with a knowledge of item writing procedures and experience in test translation and test translation review. Each translator was a specialist in either math or science. The translator of the Mathematics test had an undergraduate degree in Mathematics from a university in Paraguay. The science translator had a degree in medical anthropology from a university in Colombia. Both had experience translating standardized tests, and had previously received instruction on item writing.

Both translators were oriented to the project. The orientation included information on the MCAS program and the most frequent countries of origin of examinees who would take the MCAS in Spanish. Subsequently, the translators began work on the first draft. Their first draft was reviewed by the translation manager at SLTI, who made initial decisions about how to handle wording common to both tests, such as that found in the instructions, headers, footers, item stems, etc. He then sent each translator's work to the other with instructions that the translation be evaluated by comparing it line by line and item by item with the English version. The comments of each reviewer were sent to SLTI, where they were reviewed, and then forwarded to the original translator with further observations or recommendations.

In the case of this tryout administration, the MDOE felt it was not necessary to obtain a community review of the Spanish used on the test by local Hispanics prior to administration of the pretest. This was because the translation had been extensively reviewed already, and because by the time the Spanish version was ready for a community review, the MDOE had decided to elicit systematic feedback from teachers and students on the Spanish version following its administration. The feedback elicited from teachers concerning Spanish usage in the math

and science tests showed that they felt the Spanish version accurately reflected the English original.

Test Booklet Format

There are two basic options when administering a translation or adaptation of a test in another language. One is to produce test booklets in both languages and then allow building personnel to determine which booklet should be used by the examinee. Another is to produce the test booklet in a format that uses parallel presentations: for example, putting the two languages on facing pages. Other variations are also possible, such as putting the English on the top of the page and the Spanish on the bottom, putting the left column in English and the right column in Spanish, or having students take the test in both languages and giving the student the higher of the two scores. Related to the issue of format is whether additional time should be allotted to examinees who take the test in the bilingual format, since this dual presentation may require more time for the examinee to process.

In order to investigate these options prior to making a decision, a CD-ROM search of the ERIC database was conducted using descriptors relating to testing, bilingualism, language processing, and reading skills. While this produced a considerable number of abstracts, the search did not produce any articles that dealt directly with the subject of the format of the test booklet. Several potentially relevant articles were scanned, but no direct information on the matter was found.

In addition, a request for input was transmitted to 250 second language testing specialists who subscribe to a listserve. Over half of these subscribers live and work outside of the United States. The request for input produced several interesting comments and descriptions of local practice in different parts of the world. These were presented in a 15-page report to the MDOE (Stansfield, 1997).

The report drew three conclusions:

1. When the bilingual format has been tried, people seem to be satisfied with it.

2. When the bilingual format is not used, people cite well-intentioned and possibly valid reasons why it should not be. These reasons relate to the lack of authenticity in the bilingual test format (that is, students are not typically presented with bilingual texts to process), the fear that examinees would respond differentially to the bilingual format, and the fear that scores would not be the same as they would if the test were taken in a single language.

3. When students are given test booklets in both languages and asked to choose which booklet they would like to use, their decision may be affected by pressure from peers, teachers, or society to take the test in the societally dominant language (in this case English).

Because the investigation of test booklet formats indicated theoretical disagreement, yet user satisfaction with the various alternatives, it was decided to print the Spanish language test booklets in two formats. One format would be a Spanish only test booklet, and the other would be a bilingual or Spanish/English booklet, with the Spanish and English text on facing pages. In both cases, the classroom teacher would determine if an LEP student would get a Spanish only or bilingual version, based on the teacher's knowledge of the student's degree of literacy in Spanish and English.

It is noteworthy that the 1996 administration of the National Assessment of Educational Progress (NAEP) used both a bilingual and Spanish only version of the NAEP math exam, but an analysis of which format worked best was not available at the time a decision on test booklets had to be made. (Olsen and Goldstein, 1996)

Debriefing of Teachers and Examinees

In order to collect reactions to the different test booklet formats and a variety of other issues relating to the Spanish version, MDOE staff interviewed teachers and students after they had taken the Spanish and bilingual versions of the test. The interviews were conducted by three MDOE employees, two native Spanish speakers, and a non-Hispanic who speaks Spanish and formerly taught in a bilingual classroom. The interviewers interviewed 97 students in 19 schools in seven school districts. The sample included 34 students at grade 4, 32 at grade 8, and 31 and grade 10. Students were selected to be interviewed by the school and in most cases were interviewed individually. In addition, 17 teachers of these students were individually interviewed after they administered the tests.

Although not highly structured, the interviews produced the desired reactions to the two formats. Most students indicated that they preferred the bilingual format. Although most students who received bilingual test booklets relied mostly on the Spanish version of the items; in some cases they also read the English version and felt that they gleaned some additional meaning from this version. As a result, MDOE staff decided that in the future, only a bilingual version of the test would be printed in addition to the English version.

The interviews with students and teachers also indicated that while some students were not familiar with academic Spanish, they were satisfied with the Spanish used on the test. Teacher/administrators who had difficulty doing a sight translation of the general directions to the test, requested a Spanish translation of the directions to test administrators. Teacher/administrators also identified some printing errors, and suggested the need for a final Spanish editing of the blueline test booklet. They also made specific recommendations concerning the test administration procedures, the exemption/exclusion criteria, and the need for study materials in Spanish.

Trial Scoring of the Spanish Language Version

Since the MCAS assessments also contain open-response tasks, it was necessary to locate individuals who could score the responses written in Spanish to the open-response items. In

order for scores on achievement tests in specific subjects to be perceived as valid, those who score them must have credibility as persons who are in a position to distinguish a good performance from a poor one. This means that scorers of a math or science test should be proficient in math or science. However, there is a shortage bilingual certified math or science teachers in Massachusetts.

Because of the shortage of bilingual raters, in spring 1997, the MDOE, ASME, and SLTI devised a procedure called "consensus scoring." Consensus scoring pairs two raters who review each test performance and jointly decide on a rating. In this case, the pair consisted of a certified math or science teacher and a bilingual teacher or individual. The certified teachers had all served previously as raters of responses in English. Consensus scoring of Spanish responses began in August 1997 with a review of the test and the purposes of the item tryouts and the administration of Spanish forms. Then, scorers underwent training on the rubrics developed for assigning points to each response. Benchmark papers in English were reviewed and discussed by the raters. Then a table full of readers scored several papers written in Spanish. After agreeing on the appropriate ratings, the raters began to score papers in pairs. The bilingual individual translated the response for the certified teacher, and they jointly discussed the performance and agreed on a rating. Scorers were requested to identify potential exemplar papers for each score level that could be used to train raters of Spanish responses next year should the items appear in the operational tests.

By the end of the first day of scoring, a bilingual individual was generally able to score a paper in Spanish alone, as long as he or she had immediate access to the certified math or science teacher who was scoring in the next chair. Similarly, if the certified teacher had studied Spanish in high school or college, by the end of the day he or she had learned to recognize the critical elements of the Spanish response. As a result, the certified teacher could also generally score alone, as long as there was immediate access to the bilingual individual in the next chair. This means that the process of scoring papers written in Spanish is not twice as costly as the scoring of papers in English, as we originally had feared. Also, the consensus scoring procedure builds the pool of people who can score papers in Spanish, and it increases the number of bilingual teachers who can score papers in specific content areas, such as math and science.

At the end of the day, the scorers who had previously scored papers in English discussed the test items, the scoring process, and the responses they had read. Scorers recommended that in the future, the Reference Sheet containing mathematical formulas and other information be translated to Spanish. This is because formulas in Spanish often use different letters than in English, since words such as "width" begin with a different letter in Spanish. Scorers observed that few students performed well on the tasks in English, and even fewer performed well in Spanish. Many students did not write anything at all in Spanish, even though they answered the multiple-choice items. Nonetheless, there was general agreement in the observation that the tasks functioned in Spanish in much the same way they functioned in English. That is, they elicited responses of the same nature and structure, and these responses fit nicely to the rubrics developed for the English version. This suggests that the validity of these open-response items was not compromised by their translation to Spanish or their scoring in Spanish. Finally, the scorers agreed that the availability of a Spanish version of the instruments allowed a good number of Spanish dominant students to show what they know and can do.

Results of Item Tryouts

Table 1 shows the numbers of students taking the Spanish forms in the spring of 1997. (Spanish forms were also administered at grade 10, but the numbers of students at this grade were very few; consequently, the focus in this paper is on grades 4 and 8.) The average number of students taking an English form at each grade and in each subject ranged from 400 to 800 approximately. This information is found in Table 7. As can be seen, the Ns for each grade level are not large. This may indicate that only a fraction of Hispanic LEP students were perceived as being literate in Spanish. Further information on this issue will have to wait until data from the first operational administration (Spring 1998) is available. At this administration, the identification and testing of Spanish literate LEP students by school districts will be obligatory.

TABLE 1
Number of Students Taking Each Language Version by Grade and Subject

Grade	Subject	Form*	N
4	Mathematics	S	207
		SE	190
4	Science	S	158
		SE	171
8	Mathematics	S	97
		SE	83
8	Science	S	76
		SE	84

*S = Spanish form SE = Spanish/English bilingual form

Table 2 summarizes the performance of students on items common to the Spanish, Spanish/English (bilingual), and English forms. Items are subdivided into multiple-choice (MC) and open-response (OR) formats. The scores are grouped into one of three test booklet format categories: Spanish only test booklet (S), bilingual test booklet (SE), and English only test booklet (E).

TABLE 2

Average Total Scores* on Items Common to Spanish, Spanish/English, and English Forms

Grade	Subject	# Items	Average Total Score S	SE	E
4	Math	16 MC, 4 OR	8.89	9.01	15.36
4	Science	20 MC, 4 OR	10.36	10.93	15.93
8	Math	12 MC, 3 OR	3.19	3.48	8.27
8	Science	14 MC, 3 OR	6.80	6.48	12.95

*Sums of item means (p-values for multiple-choice)

The results in Table 2 show that the Spanish only and Spanish-English test booklets produced very similar mean scores. For three of the four groups, the bilingual format group scored slightly higher than the Spanish only group. This result agrees with the feedback obtained from students that on some items they were able to use the English version to better understand the question and select an appropriate response.

Table 2 also shows a substantial disparity in the performances of the Spanish-speaking and English-speaking students on the MCAS instruments. While such a disparity between white and Hispanic populations is consistent with the findings of other studies, including the National Assessment of Educational Progress, it is important for the purposes of developing a new test that every effort be made to assure that performance differences are not due to characteristics of the instruments that might unfairly favor the English students. Thus, the effectiveness of translation can be examined in the context of test item bias (Hambleton, 1993). Because of the limited number of students taking the Spanish forms and because the items included in these forms are not the items to be included in the operational Spanish forms, sophisticated bias analyses were not performed on the MCAS item tryout data. However, item difficulty data were generated and test items were examined in order to at least minimally examine how well the Spanish items had worked. For the operational tests, for which there will be one exactly corresponding Spanish and English form and a larger data set, more sophisticated bias analyses will be conducted, including IRT differential item functioning analyses and the examination of differential factor structures for Spanish speaking and English students.

Content Analysis of Test Items

Because the fourth grade group was the largest to take Spanish language versions of the tests, the discussion below focuses on the grade 4 tests. Tables 3 and 4 provide information on the p-values (percents answering multiple-choice items correctly) for the grade 4 Mathematics and Science and Technology tests.

As expected from the sizable differences in mean scores in favor of the English test booklet group, nearly all items were easier for students who took the test in English. This finding applies to both Mathematics and Science and Technology at all grade levels. However, the discrepancy in p-values was not uniform throughout. Naturally, the discrepancy in p-values was greater for some items than for others. In order to examine discrepancies in item difficulty for the students who took a Spanish version versus those who took an English version, the magnitude of the discrepancy was calculated. So for a given item, the p-value of students who received the English test booklet was subtracted from the p-value of students who received a Spanish test booklet. This nearly always resulted in a negative value, meaning that the p-value of the English group was greater than the p-value of the Spanish group. We also rank ordered the p-values for each group. Thus, the easiest item for a group, the one with the highest p-value, received rank 1, and the hardest, rank 20. We then identified items whose discrepancy in item difficulty and rank order was greatest and examined both the Spanish and English version of these items. For nearly all items, the examination failed to explain the reason for the difference in expected performance on that item. Indeed, most such item level examinations failed to confirm commonly held beliefs about the causes of differential performance. The examples that follow illustrate this phenomenon.

TABLE 3

P-Value Differences and Rank Order of P-Values for Multiple-Choice Items Common to Spanish, Spanish/English, and English Forms

Grade 4 Mathematics

Item #	P-Values			P-Value Differences		P-Value Rank Order		
	P_S	P_{SE}	P_E	P_S-P_E	P_{SE}-P_E	Rnk_S	Rnk_{SE}	Rnk_E
17	.68	.69	.80	-.12	-.11	1	1	1
1	.51	.52	.77	-.26	-.25	2	4	4
19	.47	.50	.79	-.32	-.29	3	5	3
21	.47	.53	.65	-.18	-.12	3	2	6
4	.44	.53	.80	-.36	-.27	5	2	1
2	.41	.32	.54	-.13	-.22	6	7	9
13	.37	.45	.54	-.17	-.09	7	6	9
7	.35	.32	.73	-.38	-.41	8	7	5
18	.31	.25	.41	-.10	-.16	9	13	11
20	.29	.31	.57	-.28	-.26	10	10	8
5	.28	.32	.35	-.07	-.03	11	7	14
3	.23	.23	.40	-.17	-.17	12	14	12
8	.22	.22	.31	-.09	-.09	13	15	15
16	.20	.27	.38	-.18	-.11	14	11	13
14	.19	.27	.65	-.46	-.38	15	11	6
15	.10	.10	.28	-.18	-.18	16	16	16

Mathematics Assessment. As can be seen in Table 3, item 4 was the easiest of the multiple choice items for the English group but several ranks lower for the Spanish groups. It also showed a discrepancy in p-value considerably larger than average. The item asks the examinee to identify which of four visuals correctly represents 7/8. Inspection of the visuals gave no clue as to the performance differential.

Item 5 showed the smallest discrepancy in p-value across the groups. This item was comparatively easier for Hispanics than the other items on the Mathematics assessment. Examination of the item showed that it tests knowledge of metric units of measure, a concept with which many Hispanics are more likely to be familiar, since the metric system is used in most Hispanic countries to one degree or another. This was the only metric item among the 20 on the math test. This finding suggests that item level performance reflects prior instruction or informal exposure to the knowledge, skill or ability (KSA) tested by the item.

Item 7 is another item that produced a large discrepancy in item difficulty. Hispanics performed comparatively worse on this item than on others. Examination of the item, which involves proportions within a circle, gives no clue as to why it was more difficult.

Science and Technology Assessment. Table 4 shows the performance on the Science and Technology assessment. None of the items on this instrument involved metric units of measure. Item 7 was the most discrepant item in terms of p-values. It had a p-value of .67 for the English group, yet for the Spanish groups its p-value was at the chance level. Examination of the item revealed that it was the one item on the test that is contextualized in a Latin American country, in this case Costa Rica. The belief that the locus of the contextualization influences performance is not supported by the results on this item.

TABLE 4

P-Value Differences and Rank Order of P-Values for Multiple-choice Items Common to Spanish, Spanish/English, and English Forms Grade 4 Science and Technology

Mult. Ch.	P-Values			P-Value Differences		P-Value Rank Order		
Item #	P_S	P_{SE}	P_E	P_S-P_E	P_{SE}-P_E	Rnk_S	Rnk_{SE}	Rnk_E
9	.73	.71	.75	-.02	-.04	1	1	2
1	.59	.56	.73	-.14	-.17	2	3	3
8	.58	.68	.54	.04	.14	3	2	9
18	.47	.42	.45	.02	-.03	4	7	14
2	.43	.44	.65	-.22	-.21	5	4	6
13	.40	.44	.80	-.40	-.36	6	4	1
6	.38	.40	.69	-.31	-.29	7	9	4
3	.37	.42	.24	.13	.18	8	7	20
17	.37	.37	.57	-.20	-.20	8	10	8
14	.35	.31	.35	0	-.04	10	15	16
22	.32	.36	.53	-.21	-.17	11	11	10
4	.31	.43	.47	-.16	-.04	12	6	13
5	.28	.32	.52	-.24	-.20	13	13	11
20	.28	.35	.64	-.36	-.29	13	12	7
15	.27	.32	.25	.02	.07	15	13	19
19	.27	.27	.39	-.12	-.12	15	16	15
7	.25	.27	.67	-.42	-.40	17	16	5
16	.23	.22	.49	-.26	-.27	18	18	12
21	.22	.22	.31	-.09	-.09	19	18	17
10	.20	.14	.26	-.06	-.12	20	20	18

Item 18, dealing with the kind of traits that can be inherited from parents, was equally difficult for the Spanish and English language examinees. There was no performance differential on this item. Yet this item stands out in that it uses a low frequency word in Spanish ("progenitores") to translate a high frequency word in English ("parent"). Several teachers even complained that the use of the word on the Spanish language version of the test was inappropriate and would make the item more difficult. Nonetheless, the item was far easier than other items for Hispanic students. The low frequency word was used because the context required it in Spanish, and the word was defined in parentheses. Nevertheless, the definition does not explain why the item was comparatively easier for Hispanic students.

Item 14, dealing with phases of the moon, functioned like item 18 in that it showed no performance differential across groups, but inspection of the item revealed nothing as to why is was easier for Hispanics than other items.

The content examination of multiple-choice items on the two language versions of the test in most cases failed to attribute the differences in difficulty of items to some inherent cultural or linguistic bias. As a result, one can accept the differences as representing differences in knowledge of the subjects assessed by the items for the groups that took the Spanish and English versions.

Results on Open-Response Items

Table 5 shows the performance differences between Spanish speaking and English students on open-response items administered in the 1997 tryouts. As with multiple-choice items, the disparity between the performances of the two groups of students is substantial.

It is difficult to judge the differences in the open-response items, since there are only four per assessment, and because the meaning of the scores is influenced by the training and the exemplars for each item. Still, the Spanish and English versions of the items were examined in order to gain an understanding of the results. The examination of the two language versions failed to contribute to an understanding of the relative performance differentials.

TABLE 5

MEAN SCORES* ON OPEN-RESPONSE ITEMS FOR SPANISH, SPANISH/ENGLISH, AND ENGLISH FORMS

Grade	Subject	Open-Response Item #	Mn_S	Mn_{SE}	Mn_E
4	Math	11	1.00	.89	1.49
4	Math	12	1.32	1.33	2.74
4	Math	23	.44	.46	.76
4	Math	24	.61	.50	1.40
8	Math	11	.16	.23	1.28
8	Math	12	.46	.41	1.36
8	Math	23	.20	.23	1.01
8	Math	24	.33	.22	-**
4	Science	11	.61	.81	1.17
4	Science	12	1.17	1.14	1.78
4	Science	23	1.13	1.22	1.85
4	Science	24	.15	.11	.83
8	Science	11	.32	.24	1.36
8	Science	12	.22	.24	-**
8	Science	23	.46	.50	1.31
8	Science	24	1.45	1.18	1.92

*Responses were scored 0,1,2,3,or 4 points.

** An item comparable to the Spanish item was not administered in any English form.

Test Reliability

The joint Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985) call for the provision of evidence of reliability on dual language versions of tests. Tables 6 and 7 show summary statistics for the Spanish and English forms administered in the spring of 1997. Recall, there are no English forms corresponding exactly to the Spanish forms in terms of the items included in the instruments. However, all forms represented in these tables included 20 multiple-choice and four open-response items.

TABLE 6

SPANISH FORM SUMMARY STATISTICS

Grade	Subject	Form*	N	Mean	Std. Dev.	Alpha
4	Mathematics	S	207	10.6	5.85	0.76
		SE	190	10.8	6.20	0.80
4	Science	S	158	10.3	3.76	0.56
		SE	171	10.9	3.70	0.48
8	Mathematics	S	97	5.47	2.85	0.48
		SE	83	5.78	3.06	0.57
8	Science	S	76	8.59	3.73	0.60
		SE	84	8.45	3.47	0.36

*S = Spanish form SE = Spanish/English form

TABLE 7

ENGLISH FORM SUMMARY STATISTICS*

Grade	Subject	N	Mean	Std. Dev.	Alpha
4	Mathematics (24 forms)	513	16.3	6.45	.79
4	Science (15 forms)	804	15.3	5.37	.75
8	Mathematics (24 forms)	398	12.8	7.21	.83
8	Science (19 forms)	511	15.6	5.70	.77

*Reported statistics are averages across forms

On the surface, the data suggest that the reliability of the Spanish forms is less than that of the English forms. However, the standard deviations reported in the tables indicate that the Spanish-speaking students represent a more homogeneous group of students than the English speaking students in terms of achievement. That is, there appears to be a restricted range of achievement represented by the data from the Spanish students.

Table 8 shows reliability coefficients for the Spanish forms adjusted for restriction of range by the following formula:

$$r'_{xx} = \sqrt{1 - \frac{s_1^2}{s_2^2}(1 - r_{xx}^2)}$$

where s_1^2 is the variance of the sample on which the original r_{xx} is based, and s_2^2 is the variance on which the adjusted r'_{xx} is to be based.

TABLE 8

Alpha Reliability Coefficients

Grade	Subject	Form*	Alpha	Adjusted Alpha**
4	Math	S	.76	.81
		SE	.80	.82
		E	.79	-
4	Science	S	.56	.81
		SE	.48	.80
		E	.75	-
8	Math	S	.48	.94
		SE	.57	.94
		E	.83	-
8	Science	S	.60	.85
		SE	.36	.82
		E	.77	-

* Alpha coefficients reported for English (E) forms are averages across 15 to 24 forms with the same numbers of items as the Spanish (S) and Spanish/English (SE) forms (20 multiple-choice and 4 open-ended).

** These alpha coefficients are adjusted for restriction of range since the distributions of Spanish student scores are considerably less variable than those of the English students.

The data in Table 8 suggest that the reliabilities of the Spanish and English forms are comparable. For grade 8 Mathematics, the adjustments in reliability coefficients appear unusually large, but it was at this grade in Mathematics that the standard deviations for scores on the Spanish forms were the smallest.

Summary of Lessons Learned

It is difficult to draw firm conclusions from the tryout data. The tryout involved small samples of Spanish speaking students who took only the Spanish language version of the test. These versions were short (20 items), while the versions used in the operational administration in May 1998 will contain 50 items. We expect a considerably larger number of Hispanic students to participate, since inclusion of all appropriate students in the testing program will be obligatory. With a larger, comprehensive sample of Spanish literate examinees and longer tests, we expect to conduct further analyses in order to better address issues of validity, item bias, and score comparability in the context of an adaptation of a standards-based state assessment.

We did learn a good deal through the trial administration and scoring of the Spanish items. The main points learned were the following:

1. Although only a fraction of Hispanic LEP students were judged by their teachers to be able to benefit from taking the test in Spanish, those identified were able to participate in the assessment. These students either could not take or would have been disadvantaged by the English only version of the assessment.

2. A bilingual test booklet, with Spanish on the left page and English on the right, is viewed more positively by students than a Spanish only booklet. While students rely mostly on the Spanish language pages, they sometimes examine the facing English page to supplement their understanding of the item. This resulted in slightly higher scores on the bilingual version of the test.

3. The consensus scoring of open-response items works well. While its cost is high, it is not as high as one might expect. Also, the consensus scoring method adds to the small pool of experienced raters who can score in both English and Spanish.

4. The iterative translation procedures followed produced a Spanish language version that was positively received by Spanish-literate examinees representing a variety of national origins.

5. Experienced raters felt that the open-response items and the scoring guide functioned in Spanish as they had functioned in English. At this point, there is no cause for alarm that the Spanish language versions will function differently from the English versions. This observation applies to the open-response items and the test as a whole.

6. The Spanish language versions seem to be as reliable as the English language versions, when adjusted for restriction of range.

7. Content analysis of items that appeared to be comparatively easier or more difficult for the Spanish groups normally did not reveal the reason for that difference. This finding was not unexpected, since items were reviewed for content sensitivity and bias prior to testing.

By developing and administering a Spanish language version of the state assessments, Spanish literate students who have recently arrived in the US inure the benefits of inclusion in the state assessment system. These benefits include early diagnosis of their KSAs as elaborated in the state curriculum frameworks, and the acquisition of an experience base that they can apply when they later have to take the tests in English. Furthermore, the Spanish language version allows the state and each district to increase the percentage of LEP students included in the assessment, and to disaggregate the scores of this group. Thus, it appears that the Spanish language versions will help address the assessment needs of students, teachers, and parents, as well as the accountability needs of state and district officials.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Hambleton, R.K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment, 9* (1), 57-68.

Lacelle-Peterson, M. & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64*(1), 55-75.

Massachusetts Department of Education. (January 1998). *Guide to the Massachusetts Comprehensive Assessment System: Science and Technology*. Malden, MA: Massachusetts Department of Education.

Olsen, J. & Goldstein, A. (July, 1996). *Increasing the Inclusion of students with disabilities and limited English proficiency students in NAEP*. Washington, DC. National Center for Education Statistics, US Department of Education.

Stansfield, C.W. (September, 1996). *Content Assessment in the Native Language*. ERIC Digest. Washington, DC: ERIC Clearinghouse on Assessment and Evaluation.

Stansfield, C.W. (January, 1997). *Experiences and Issues Related to the Format of Bilingual Tests: Dual Language Test Booklets versus Two Different Test Booklets*. Report to the Massachusetts Department of Education. N. Bethesda, MD: Second Language Testing, Inc.

Spanish/English

Grade 4

Massachusetts Comprehensive Assessment System

Spring 1997 Question Tryout

Mathematics

Student Name:_____

School Name:_____

District Name:_____

Massachusetts Department of Education

GENERAL DIRECTIONS

You will be answering questions in either mathematics or science and technology. This question tryout is divided into two sessions. The first session contains ten multiple-choice questions and two open-response questions. The second session contains multiple-choice, short-answer, and open-response questions. You will have as much time as you need to complete each session.

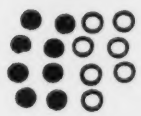
Note:

The MCAS question tryout forms are secure material. They may not be duplicated in any way. All MCAS question tryout materials must be returned to Advanced Systems after the administration is complete.

MASSACHUSETTS DEPARTMENT OF EDUCATION
MCAS
MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM

4. ¿Cuál de las respuestas representa correctamente $\frac{7}{8}$?

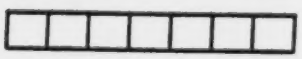
Andre

A. 

Benita

B.

Corey

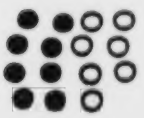
C. 

Roselaure

D.

4. Which picture correctly shows $\frac{7}{8}$?

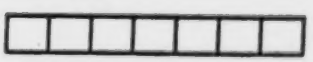
Andre

A. 

Benita

B.

Corey

C. 

Roselaure

D.

SÍRVASE PASAR A LA PÁGINA SIGUIENTE→

NO SE PUEDEN DUPLICAR.

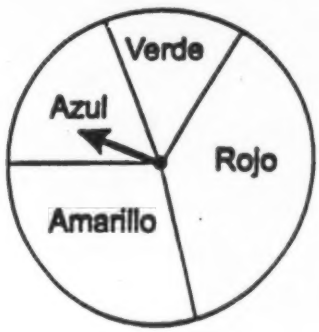
PLEASE GO ON TO THE NEXT PAGE→

NOT BE DUPLICATED.

Emplee la ilustración siguiente para contestar la pregunta número 5:

5. La mejor unidad para expresar el peso de una grapa de papel es
   A. centímetros.
   B. litros.
   C. kilogramos.
   D. gramos.

7. Este es un trompo para un juego.

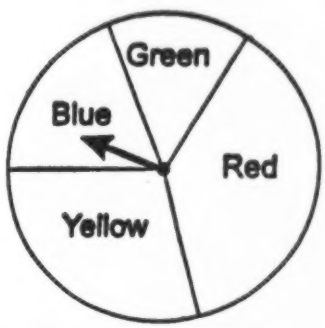


¿Qué color es más probable que caiga?
A. azul
B. verde
C. amarillo
D. rojo

Use the illustration below to answer question 5.

5. The best unit to use to weigh a paper clip is
   A. centimeters.
   B. liters.
   C. kilograms.
   D. grams.

7. This is a spinner for a game.



Which color are you most likely to spin?
A. blue
B. green
C. yellow
D. red

MATERIALES CONFIDENCIALES. NO SE PUEDEN DUPLICAR.

7. El Monte Arenal, un volcán de Costa Rica, está en erupción. Emite calor durante la erupción. ¿Cuál es la fuente de calor?

   A. El calor proviene del sol.
   B. El calor proviene de lagunas calientes de agua subterránea.
   C. El calor proviene del centro de la Tierra.
   D. El calor proviene de las plantas y los animales en descomposición.

7. Mount Arenal, a volcano in Costa Rica, is erupting. Heat is being released during the eruption. What is the source of the heat?

   A. The heat comes from the sun.
   B. The heat comes from pools of underground water.
   C. The heat comes from the center of Earth.
   D. The heat comes from decaying plants and animals.

BEST COPY AVAILABLE

MATERIALES CONFIDENCIALES. NO SE PUEDEN DUPLICAR.

18. ¿Cuál es una característica que un perro NO hereda de sus progenitores [los animales que le dieron la vida]?

A. El largo del pelo.
B. La forma de la nariz.
C. El apetito.
D. El color del pelo.

18. Which of the following is a trait that a dog does NOT inherit from its parents?

A. the length of its fur
B. the shape of its nose
C. the size of its appetite
D. the color of its fur

BEST COPY AVAILABLE

MATERIALES CONFIDENCIALES. NO SE PUEDEN DUPLICAR.